AIChE WILEY InterScience DISCOVER SOMETHING GREAT

Prediction of Properties in Homologous Series with a Shortcut QS2PR Method

Georgi St. Cholakov
Dept. of Organic Synthesis and Fuels, University of Chemical Technology and Metallurgy, Sofia 1756, Bulgaria

Roumiana P. Stateva
Institute of Chemical Engineering, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria

Mordechai Shacham
Dept. of Chemical Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel

Neima Brauner
School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel

DOI 10.1002/aic.11059

For prediction of properties of pure compounds in homologous series a shortcut version of the authors' QS2PR technique is introduced. The structure–structure relationship between a target compound and three predictive compounds from its homologous series is described in the short version with the number of carbon atoms and one easily calculated noncollinear molecular descriptor. This correlation provides predictions of high precision for properties of members of the homologous series, if property data for interpolation are available. It can also be used for short-range extrapolation with precision similar to that of asymptotic relationships. The advantages of the proposed method over the existing methods are that with the structural correlation parameters it predicts all structure-dependent properties of the target and enables estimation of its missing properties even if properties for only three predictive compounds are available. The method can also use as a descriptor the normal boiling temperature. © 2006 American Institute of Chemical Engineers *AIChE J,* 53: 150–159, 2007

Keywords: property prediction, QSPR, molecular descriptors, homologous series

Introduction

Pure component properties are essential to chemical engineering calculations in such processes as steady-state and dynamic simulation, process and product design, environmental impact assessment, and hazard and operability analysis.[1] However, published experimental data are scarce and often contradictory. Moreover, there are compounds for which some properties cannot be measured (such as when the compound decomposes before reaching its critical properties), and extrapolated values must be used. Therefore, it is very important to develop reliable prediction techniques for such compounds.

Correlations using only the number of carbon atoms for predicting particular properties of compounds of some homologous series have been suggested by many authors.[2,3] They are usually based on the Flory–Huggins "mers" approach. Because the carbon number–property relationships

---

Correspondence concerning this article should be addressed to M. Shacham, at shacham@bgumail.bgu.ac.il.

are asymptotic such correlations are often called ABCs (asymptotic behavior correlations). A review of the state of the art, including reference to the latest experimental data and a new method, developed by the authors, was recently published by a well-recognized team.[4]

In the above approach, a different ABC has to be derived for each property. Typically, all available experimental property data are used to determine the several constants in each correlation. ABCs provide very high precision when used for interpolation or for extrapolation close to the member with the highest number of carbon atoms for which property data are available.[5] However, when extrapolating far from the available data, the discrepancies of property values predicted by different authors can be very significant.[3,6,7] ABCs also typically do not include some of the first members of the series for which the influence of the end groups is very strong. To overcome the problems outlined, the correlations usually use a starting property value and a limiting property value at $n \to \infty$, together with two or more adjustable parameters characterizing the asymptotic influence of the repeating groups.[2,3] Group contribution methods assigning physical meaning to some of the parameters have also been proposed for homologous series.[4,8,9]

In any case all parameters have to be determined from experimental data. It can be expected that the accuracy of the prediction (such as that of critical properties obtained for an unstable target compound) deteriorates, because the measured data for its neighboring compounds are less reliable, than for the remote stable compounds, in the same homologous series. Typically different authors choose different values for compounds with $n \to \infty$, which affects the predicted values. Alternatively, prediction of standard liquid density (such as at 20°C) becomes hypothetical for compounds, which are solid at the standard temperature. Because of experimental problems the property data that can be measured are naturally limited. For instance, only recently have data for the critical properties of *n*-alkylcyclohexanes up to *n*-butylhexane been obtained.[10] Thus, deviations of long-range predictions cannot really be judged.

Despite the difficulties, prediction with adequate precision especially for compounds with higher numbers of carbon atoms remains important for both theoretical and practical reasons.[4,9,11]

Recently, Shacham et al.[12] and Brauner et al.[13] demonstrated that most pure component properties of a target compound can be represented, with high precision, as a linear combination of the same properties of structurally similar predictive compounds. The structurally similar predictive compounds are identified by a structure–structure correlation, derived from calculated molecular descriptor data for the target and the potential predictive compounds. Structure-dependent properties of the target compound can then be predicted by replacing the molecular descriptor values of the predictive compounds [on the right-hand side (rhs) of the structure–structure equation] by the corresponding property values. Thereby, a property–property correlation is obtained by a method denoted as QS2PR (quantitative structure–structure–property relationship).

The development of QS2PRs requires access to a large database of molecular descriptors. In our previous studies, we used a descriptor database with 99 molecular descriptors for 260 hydrocarbons and data for the normal boiling temperatures, the critical properties, and the liquid densities of the compounds[5,11] and derived the structure–structure correlations with a stepwise regression algorithm (the SROV program) of Shacham and Brauner.[14] Details about the database were presented in our previous publications (see, for example, Brauner et al.[15]). However, it seems reasonable to expect that the number of the necessary predictive compounds and descriptors is determined by the structural variation of the target compounds and the range of extrapolation required.

Consequently, the potential for simplifying the QS2PR technique for members of homologous series in which the structural variation is naturally limited is explored in this article. The DIPPR[16] and the NIST[17] databases are used as sources of published (experimental or predicted) property data.

Principles of the QS2PR Technique

Because the QS2PR technique was previously described in detail,[12] here we provide only a brief outline of its main features as an introduction to its shortcut application to members of homologous series.

Let us assume that the vector of properties of the target compound $\mathbf{y}$ (the dependent variable) is potentially related to a set of m vectors of properties of predictive compounds (independent variables) $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. The following partition of the $\mathbf{y}$ and $\mathbf{x}$ vectors to subvectors is used:

$$\mathbf{y} = \begin{Bmatrix} \mathbf{y}_c \\ \mathbf{y}_p \end{Bmatrix} \qquad \mathbf{x}_i = \begin{Bmatrix} \mathbf{x}_{cj} \\ \mathbf{x}_{pj} \end{Bmatrix} \tag{1}$$

where $\mathbf{y}_c$ is an N vector of known properties; $\mathbf{y}_p$ is a K vector of unknown properties. Both the N vector $\mathbf{x}_{cj}$ and the K vector $\mathbf{x}_{pj}$ contain known properties. Typically, the subvectors $\mathbf{y}_c$ and $\mathbf{x}_{cj}$ contain properties that are directly related to the molecular structure and can be calculated with high accuracy (molecular descriptors), whereas the subvectors $\mathbf{y}_p$ and $\mathbf{x}_{pj}$ contain measured properties with various levels of experimental error. We seek to model the structure–structure relationship between $\mathbf{y}_c$ and the independent variables $\mathbf{x}_{c1}, \mathbf{x}_{c2}, \ldots, \mathbf{x}_{cm}$ by a linear regression model, with the general form

$$y_{ci} = \beta_1 x_{c1i} + \beta_2 x_{c2i} \cdots \beta_m x_{cmi} + \varepsilon_i \tag{2}$$

where the weighing factors $\beta_1, \beta_2 \cdots \beta_m$ are the model parameters to be estimated and ε_i represents independent normal errors of a constant variance.

The practical application of Eq. 2 requires preparation of a bank of potential predictive compounds as a database. The same set of molecular descriptors must be defined for all compounds included in the database, whereas the span of the molecular descriptors should reflect the difference between the compounds in the database. Having the corresponding molecular descriptors for a target compound $\mathbf{y}_c$ defined as well, a stepwise regression procedure is applied to the database to identify the most appropriate predictive compounds that should be included in the structure–structure regression model (Eq. 2) and obtain the respective model parameters. Upon identifying the model parameters, the following equa-

Table 1. Selected Molecular Descriptors and Properties for Members of the *n*-Alkane Homologous Series

No.	Compound	C Atoms	Molar Mass	Wiener's Index*	Temperature (K)*		Critical Pressure (MPa)*
					Boiling	Critical	
1	*n*-Octane	8	114.232	84	398.7	568.7	2.49
2	*n*-Nonane	9	128.259	120	423.8	594.6	2.29
3	*n*-Decane	10	142.286	165	447.2	617.7	2.11
4	*n*-Undecane	11	156.313	220	468	639	1.98
5	*n*-Dodecane	12	170.340	286	489	658	1.82
6 (target)	*n*-Tridecane	13	184.367	364	507	675	1.68
7	*n*-Tetradecane	14	198.394	455	523	693	1.57
8	*n*-Pentadecane	15	212.421	560	543.78	708	1.48
9	*n*-Hexadecane	16	226.448	680	559.94	723	1.40
10	*n*-Heptadecane	17	240.475	816	575.3	736	1.34
11	*n*-Octadecane	18	254.502	969	589.86	747	1.29
12	*n*-Nonadecane	19	268.529	1140	603.76	755	1.16
13	*n*-Eicosane	20	282.556	1330	616.92	768	1.07
14	*n*-Beneicosane	21	296.583	1540	629.65	778	1.03

*From the database of Cholakov et al.[11]

tion can be used to predict unknown properties of the target compound:

$$\mathbf{y}_p = \beta_1 \mathbf{x}_{p1} + \beta_2 \mathbf{x}_{p2} \cdots \beta_m \mathbf{x}_{pm} \qquad (3)$$

The properties that can be predicted for the target compound include all properties that are available for all the predictive compounds included in the structure–structure correlation.

Application of the QS2PR Method to Homologous Series

Application of the QS2PR method to predict properties of members of homologous series, and the basis for the development of simplified correlations in this case, will be demonstrated with the *n*-alkane series. Data for selected members of this series are shown in Table 1. The data include the normal boiling temperature (T_b), the critical temperature (T_c), the critical pressure (P_c), and three molecular descriptors—number of carbon atoms, molecular mass, and the integer Wiener index—calculated as described by Labanowski et al.[18] The latter will be used for the simplified version of the QS2PR for homologous series (SC-QS2PR) as shown below.

Applying the QS2PR technique with all 99 descriptors to select predictive compounds for *n*-tridecane (compound No. 6 in Table 1, target compound) yields a large number of potential structure–structure correlations with three predictive compounds that achieve the precision needed to predict the properties of the target. The predictive compounds included in the various models, the standard deviation of the structure–structure correlation, and the percentage error in predicting T_b, T_c, and P_c of the target compound are shown in Table 2.

The errors in the predicted values of T_b and T_c are all $<1\%$, which is of the same level as the experimental error (reported in the DIPPR database[16]) for most of the compounds involved. The error in the predicted P_c seems high in cases IV and V (5.84 and 9.39%, respectively), although the experimental error reported for this property is even higher (about 10% for *n*-tridecane and 25% for *n*-eicosane, for example).

Table 2. Potential Structure–Structure Correlations and Model Precision Indicators for Representing the Properties of *n*-Tridecane (Target Compound)

No.	Compound	Predictive Compounds Included in Models No.				
		I	II	III	IV	V
1	*n*-Octane					*
2	*n*-Nonane				*	
3	*n*-Decane			*		
4	*n*-Undecane		*			
5	*n*-Dodecane	*				
6	*n*-Tridecane	target	target	target	target	target
7	*n*-Tetradecane	*				
8	*n*-Pentadecane	*	*			
9	*n*-Hexadecane			*		
10	*n*-Heptadecane		*		*	
11	*n*-Octadecane					*
12	*n*-Nonadecane			*		
13	*n*-Eicosane				*	
14	*n*-Heneicosane					*
Model precision indicators	Standard deviation	1.70E−04	1.79E−04	3.34E−04	3.67E−04	6.08E−04
	Prediction error (%) in T_b	0.69	0.79	0.42	0.48	0.5
	Prediction error (%) in T_c	0.28	0.89	0.65	0.62	0.45
	Prediction error (%) in P_c	0.58	0.51	1.99	5.84	9.39

The above example demonstrates that when the QS2PR technique is applied to property prediction of members of homologous series, the limited structural change within a homologous series allows for adequate representation of the target structure with *only* three (sometimes even two) predictive compounds. Therefore, for homologous compounds the QS2PR method can be simplified, given that only three parameters would be required in a structure–structure correlation with three predictive compounds.

The Shortcut Version of the QS2PR Method (SC-QS2PR) for Homologous Series

As shown below, the minimum information required for deriving a structure–structure relation for a target compound in terms of three predictive compounds, which are all member of the same homologous series, is the availability of two noncollinear molecular descriptors for both the predictive and the target compounds.

To find the coefficients of the structure–structure relation, $y_c = \Sigma_{j=1}^{3} \beta_j x_{cj}$ the following system of three linear equations has to be solved:

$$\begin{aligned} &\beta_1 + \beta_2 + \beta_3 = 1 \\ &\beta_1 n_{p1} + \beta_2 n_{p2} + \beta_3 n_{p3} = n_t \\ &\beta_1 x_{C1} + \beta_2 x_{C2} + \beta_3 x_{C3} = y_C \end{aligned} \tag{4}$$

where n_{p1}, n_{p2}, and n_{p3} are the numbers of carbon atoms of the predictive compounds and n_t is the number of carbon atoms of the target compound. The first expression in the Eq. 4 system reflects the influence of descriptors that have the same value for the target and predictive compounds (such as the number of CH_3 groups in the *n*-alkane series). The last expression in the Eq. 4 system must be applied with a molecular descriptor that is noncollinear with the number of carbon atoms (and is not identical for all the predictive compounds). The integer Wiener's index (*W*) with traditional presentation of bonds was chosen for the present work because it is easily calculated from the distance matrices of the molecules:

$$W = \frac{1}{2} \sum_{i,j} d_{ij} \tag{5}$$

where the compound is represented by a hydrogen depleted molecular graph; d_{ij} represents the elements of the distance matrix—the number of edges (bonds) in the shortest path between vertices (atoms) *i* and *j*; the sum runs over the number of edges between all pairs of vertices of the graph.[16] For members of homologous series simple shortcut formulae can be worked out. Regression of the *W* indices calculated by Eq. 5 for the *n*-alkanes, for instance, gives a formula that reproduces the integer values of *W* with sufficient precision:

$$W = 0.16667n^3 - 0.16667n \tag{6}$$

where *n* is the number of carbon atoms, the integer values of *W*. Equation 6 also t demonstrates hat *W* is noncollinear with the number of carbon atoms because n^3 is obviously a nonlinear term. Collinearity leads to the nearly singular or singular coefficient matrix in Eq. 4, thus preventing one from obtaining an accurate solution to this system.

The effects of collinearity (often referred to as "degree of coupling" between the parameters) can be best assessed using the condition number of the matrix of coefficients of Eq. 4. The condition number is defined as $\kappa(\mathbf{A}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ n_{p1} & n_{p2} & n_{p3} \\ x_{c1} & x_{c2} & x_{c3} \end{bmatrix}$$

The condition number provides an estimate on the upper bound of the relative errors in the calculated parameter values. Thus, $\|\Delta\beta/\beta\| \leq \kappa(\mathbf{A})\|\Delta\mathbf{x}_c/\mathbf{x}_c\|$, where $\Delta\mathbf{x}_c$ represents the noise (rounding errors) in the $\mathbf{x}_c$ values (such as the Wiener index data). A higher level of collinearity between the descriptors (higher degree of coupling) leads to larger $\kappa(A)$ values, which in turn causes a greater change of parameter values when a small change in the descriptor values is introduced. Thus, variation in the calculated parameter values can be minimized by selecting descriptors with a low level of collinearity with the number of carbon atoms (that is, low condition number) and by representing the calculated $\mathbf{x}_c$ values at high accuracy.

Noncollinearity with the number of carbon atoms can be easily verified by dividing the molecular descriptor values by the number of carbon atoms. A constant (or nearly constant) value of this ratio for the whole series indicates collinearity. For example, dividing the molecular masses in Table 1 by the respective carbon atom numbers yields a nearly constant value (between 14.279 and 14.123). In contrast, dividing the Wiener's index by the carbon atom numbers yields the series: 10.5, 13.33, 16.5, 20,. . . .

Equation 4 can be rewritten in an explicit form:

$$\begin{aligned} \beta_3 &= \frac{(y_C - x_{C1})(n_{p2} - n_{p1}) - (n_1 - n_{p1})(x_{C2} - x_{C1})}{(x_{C3} - x_{C1})(n_{p2} - n_{p1}) - (n_{p3} - n_{p1})(x_{C2} - x_{C1})} \\ \beta_2 &= -\frac{n_{p3} - n_{p1}}{n_{p2} - n_{p1}}\beta_3 + \frac{n_1 - n_{p1}}{n_{p2} - n_{p1}} \\ \beta_1 &= 1 - \beta_2 - \beta_3 \end{aligned} \tag{4a}$$

After the coefficients β_1, β_2, and β_3 have been determined by solving Eq. 4a, they are introduced into Eq. 3 to obtain the desired properties of the target compound. The maximal error in the predicted property value (Δy_p) can be estimated using the experimental errors reported for the predictive compound data (Δx_{pi}) and the error propagation formula, whereby $\Delta y_p = \Sigma_i |\beta_i| \Delta x_{pi}$.

If there are only two predictive compounds, for which experimental data are available, and they are in the near vicinity of the target, a structure–structure relation based on the number of carbon atoms only can be used. Then Eq. 4 is reduced to $\beta_1 = (n_t - n_{p2})/(n_{p1} - n_{p2})$; $\beta_2 = 1 - \beta_1$. It should be noted, however, that the use of more than three predictive compounds can be beneficial in obtaining more accurate predictions (as in long-range extrapolation). In such cases, the number of noncollinear molecular descriptors and coefficients, and thus also the number of equations, must be increased to match the number of the predictive compounds.

Table 3. Structure–Structure Correlation Data and Coefficients for Ethylcyclopentane, Heptanoic Acid, and 1-Hexanol

No.	Compound	C Atoms	Wiener's Index	Boiling Temp. (K)*	Coefficient
1	Methylcyclopentane	6	26	344.96	0.34783
2 (target)	Ethylcyclopentane	7	43	376.62	
3	Propylcyclopentane	8	67	404.11	0.95652
4	Butylcyclopentane	9	99	429.8	−0.30435
1	Hexanoic acid	6	79	478.85	0.34483
2 (target)	Heptanoic acid	7	114	496.15	
3	Octanoic acid	8	158	512.85	0.96552
4	Nonanoic acid	9	212	527.15	−0.31034
1	1-Pentanol	5	35	410.9	0.34783
2 (target)	1-Hexanol	6	56	430.44	
3	1-Heptanol	7	84	447.3	0.95652
4	1-Octanol	8	120	467.85	−0.30435

*From the database of Cholakov et al.[11]

Applying the SC-QS2PR Method for Homologous Series

Property prediction with the coefficients of the structural correlation

The SC-QS2PR method outlined above was tested with members of the *n*-alkane, 1-alkene, *n*-alkylcyclopentane, *n*-alkylbenzene, 1-*n*-alcohol, 1-*n*-alkylcarboxylic acid, 1-*n*-Cl-alkane, and 1-Br-*n*-alkane homologous series and found to be applicable to all of them. This result could be expected because from a structural perspective, in homologous series only the functional groups are different. Herein, capabilities of the SC-QS2PR method to predict numerous properties of the target, when experimental data for the predictive compounds are available, are illustrated with ethylcyclopentane, heptanoic acid, and 1-hexanol as target compounds.

The data used to derive the structure–structure correlation for the various compounds are shown in Table 3. For ethylcyclopentane, three of the neighboring compounds—methyl-, propyl- and butyl-cyclopentane—were used as predictive compounds. Introducing the numerical values of the number of carbon atoms and Wiener's index into Eq. 4 yields the following results: $\beta_1 = 0.34783$, $\beta_2 = 0.95652$, and $\beta_3 = -0.30435$. These coefficients are used in the property–property correlation to predict various properties of ethylcyclopentane. For example, introducing the boiling point data from Table 3 and the coefficients into Eq. 3 yields an estimate for the boiling temperature of ethylcyclopentane:

$$(0.34783) \times (344.96 + 0.95652) \times (404.11 - 0.30435) \times (429.8) = 375.72\ \mathrm{K}$$

Comparison of the predicted value with the experimental data shown in Table 3 (376.62 K) indicates a prediction error of 0.24%.

By introducing experimental data for other properties of the predictive compounds in the structure–structure correlation, the respective properties for the target compounds can be similarly predicted. Table 4 shows the errors in predicting 28 constant properties of the three target compounds using the so-obtained property–property relations. Out of the 84 predicted property values, the prediction error is <1% in 58 cases. In most cases, when the prediction error is >1%, it is still consistent with the experimental error level reported by DIPPR.[16] For example, the prediction error in the critical volume of heptanoic acid is 3.76%, whereas the "reliability" of these property data in the DIPPR database is <5% (meaning that the experimental error is believed to be <5%).[16] The prediction error in the lower flammability limit of 1-hexanol is 8.33%, whereas the reported reliability is <25%.

The prediction errors in the properties associated with the solid phase (melting point temperature and heat of fusion at the melting point) are inconsistent with the reported experimental errors. It should be noted that similar inconsistency exists, with respect to those properties, when the full-scale QS2PR method is used (see Brauner et al.[15]). These properties require specific descriptors, which at present are not available in the authors' database to describe the different packing of the molecules with odd and even numbers. A comprehensive review and explanation of the problems with predicting properties of the solid phase, complemented with originally developed suitable descriptors and adequate models, can be found in a recent publication.[19]

Influence of the distance from the target on prediction quality

Calculation of the various QS2PR parameters is carried out using the shortcut QS2PR. That is, only the number of carbon atoms and Wiener's index are used, instead of the vector of the 99 descriptors used earlier. The resultant parameter values are shown in Table 5.

The prediction errors (obtained by comparing reported and predicted values) for the various models and properties using the parameter values of Table 5 are shown in Table 6. The data for the properties of all compounds are experimental. The prediction errors are similar (and even smaller) when compared to those obtained using the full version of the QS2PR technique (see Table 2). The excessive prediction errors for P_c in cases IV and V can be explained by the excessive experimental error level for this property (such as a level of 10% for *n*-tridecane and 25% for *n*-eicosane). Thus, it can be concluded that the SC-QS2PR method yields prediction errors within experimental error level irrespective of the distance between the target and the predictive compounds when interpolation is used. Although similar precision can be obtained with asymptotic correlations of the carbon number, one obvious advantage of the SC-QS2PR method is that the coefficients of the structural correlation are used to predict

Table 4. Percentage Error in Predicting Properties of Ethylcyclopentane, Heptanoic Acid, and 1-Hexanol

			Ethylcyclopentane		Heptanoic Acid		1-Hexanol	
No.	Property	Unit	Reported Value*	Prediction Error (%)	Reported Value*	Prediction Error (%)	Reported Value*	Prediction Error (%)
1	Critical temperature	K	569.5	0.55	677.3	0.02	610.3	0.13
2	Critical pressure	Pa	3,400,000	0.61	3,040,000	0.22	3,420,000	0.51
3	Critical volume	m^3 kmol^{-1}	0.375	0.44	0.468	3.76	0.387	2.27
4	Crit compress factor	Unitless	0.269	0.63	0.253	3.67	0.261	2.25
5	Melting point temp.	K	134.71	7.07	265.83	6.79	228.55	4.46
6	Triple-Pt temperature	K	134.71	7.07	265.83	6.80	228.55	4.46
7	Normal boiling temp.	K	376.62	0.24	496.15	0.01	430.55	0.06
8	Liq molar volume	m^3 kmol^{-1}	0.128748	0.22	0.142369	0.54	0.1252	0.02
9	IG heat of formation	J kmol^{-1}	−1.27E+08	0.24	−5.36E+08	0.31	−3.17E+08	0.01
10	IG Gibbs of formation	J kmol^{-1}	4.48E+07	0.06	−3.34E+08	0.61	−1.34E+08	0.78
11	IG absolute entropy	J kmol^{-1} K	3.78E+05	0.25	4.80E+05	0.14	4.40E+05	0.09
12	Std heat of formation	J kmol^{-1}	−1.63E+08	0.40	−6.11E+08	0.00	−3.78E+08	0.01
13	Std Gibbs of formation	J kmol^{-1}	3.76E+07	0.45	−3.6.+08	0.01	−1.50E+08	2.12
14	Std absolute entropy	J kmol^{-1} K	2.80E+05	0.34	3.24E+05	0.20	2.87E+05	2.71
15	Heat fusion at melt Pt	J kmol^{-1}	6.87E+06	24.26	1.54E+07	28.75	1.54E+07	9.31
16	Std net heat of comb	J kmol^{-1}	−4.28E+09	0.00	−3.84E+09	0.00	−3.68E+09	0.09
17	Acentric factor	Unitless	0.270095	3.23	0.756364	0.04	0.576355	2.09
18	Radius of gyration	M	3.73E−10	0.84	4.84E−10	1.46	4.14E−10	3.10
19	Solubility parameter	$(J \cdot m^{-3})^{0.5}$	16300	0.03	22700	1.99	21800	0.92
20	van der Waals volume	m^3 kmol^{-1}	0.0704	0.04	0.0843	0.08	0.0729	0.05
21	van der Waals area	m^2	8.87E+08	0.32	1.19E+09	0.55	1.03E+09	0.57
22	Refractive index	Unitless	1.4173	0.07	1.421	0.00	1.4161	0.03
23	Flash point	K	269	0.29	382	0.61	333	0.10
24	Lower flammability limit	vol % in air	1.1	1.58	1.1	3.14	1.2	8.33
25	Upper flammability limit	vol % in air	6.7	7.92	7.2	0.33	8.2	2.65
26	Lower flamm. limit temp.	K	270	0.29	381	0.60	333	0.10
27	Upper flamm. limit temp.	K	303	0.20	420	0.15	366	0.00
28	Auto ignition temp.	K	533.15	6.68	633	0.50	558	0.59

*Data from the DIPPR database.[16]

Table 5. Structure–Structure Correlation Parameters by Eq. 4 for *n*-Tridecane (Target Compound)

		Coefficients of Models No.				
No.	Compound	I	II	III	IV	V
1	*n*-Octane					0.340426
2	*n*-Nonane				0.34585	
3	*n*-Decane			0.355556		
4	*n*-Undecane		0.348837			
5	*n*-Dodecane	0.341463				
6	*n*-Tridecane	target	target	target	target	target
7	*n*-Tetradecane	0.97561				
8	*n*-Pentadecane	−0.31707	0.953488			
9	*n*-Hexadecane			0.933333		
10	*n*-Heptadecane		−0.30233		1.065217	
11	*n*-Octadecane					1.191489
12	*n*-Nonadecane			−0.28889		
13	*n*-Eicosane				−0.41107	
14	*n*-Heneicosane					−0.53191

all structure-dependent properties. Because the influence of the distance between the target and the predictive compounds in the homologous series is not significant, prediction by interpolation can be done even when experimental data for only three predictive compounds in the series are known, which is an additional advantage.

In the SC-QS2PR technique predictions are obtained by summation instead of minimization of the experimental errors of the predictive compounds (as in traditional QSPRs). However, error propagation analysis shows that this does not significantly affect the prediction results when interpolating.

If we assume that the experimental error is about the same for all the predictive compounds (say Δx_p) the propagated error in the predicted property values of the target compound can be estimated from the error propagation formula: $\Delta y_p = \Delta x_p \, \Sigma_i \, |\beta_i|$. Using the coefficient values in Table 5 the error multiplier $\Sigma_i \, |\beta_i|$ can be calculated. It can be seen that the value of this multiplier for interpolation is between 1.2 and 2.0 and it increases slightly with increasing distance of the predictive compounds from the target compound.

Use of measured properties as descriptors

The SC-QS2PR technique can be applied by using measured properties instead of the calculated molecular descriptor (such as Wiener index). This option is useful because an experimental value for the normal boiling temperature is available for many compounds.

Prediction of the properties of ethylcyclopentane, heptanoic acid, and 1-hexanol (shown in Table 4) was repeated using the normal boiling temperatures of the predictive and the target compounds instead of the Wiener index in Eq. 4a. The experimental normal boiling temperatures shown in Table 3 were used; the coefficients obtained for the various predictive compounds were different from those shown in Table 3. For example, the structure–structure relation obtained for heptanoic acid as a target is heptanoic = (0.44444 × hexanoic) + (0.66667 × octanoic) − (0.11111 × nonanoic). However, the predicted property values and the prediction errors were not significantly different from those obtained by using the Wiener index (shown in Table 4), and with no clear trend that may indicate whether the use of the boiling temperature or the Wiener index is more advantageous.

To further demonstrate the technique, properties of three compounds from different homologous series were predicted using the normal boiling temperature in the derivation of the structure–structure correlation. The data used to derive the correlations and the resultant coefficient values for the various compounds are shown in Table 7. The target compounds are butylcyclopentane, 1-chlorohexane, and 1-bromobutane. It should be noted that in two of the cases the prediction involves extrapolation (the predictive compounds for the target butylcyclopentane are methyl-, ethyl-, and propylcyclopentane; for the target 1-chlorohexane the predictive compounds are 1-chloropropane, 1-chlorobutane, and 1-chloropentane).

Table 8 shows the prediction results for 23 constant properties of the butylcyclopentane using the so-obtained property–property relations. The table shows the property values reported in the DIPPR database[16] and the assigned reliability, the "Data type" (predicted experimental or defined), the value predicted using the SC-QS2PR method, and the prediction error. It can be seen that the prediction error is smaller (or much smaller) than the "Reliability" of the data. (Note that for the acentric factor no reliability is assigned because it is a defined property.)

Table 9 shows the predicted critical properties and liquid molar volume at 25°C for 1-chlorobenzene and 1-bromobutane using the predictive compounds and property–property correlation coefficients shown in Table 7. For 1-chlorohexane experimental values are available only for T_b and T_c in the NIST[17] database, whereas for the predictive compounds all data are available (experimental or predicted) in the DIPPR database.[16] Comparison of the predicted T_c value with the experimental value shows excellent agreement. There are no experimental data for P_c, but the predicted T_c/P_c ratio (19.86) verifies well against the same ratio calculated from UNIFAC parameters (19.28), as suggested by Zbogar et al.[20]

Table 6. Errors in Predicting the Properties of *n*-Tridecane Using the Model Parameters of Table 5

	Percentage Error in Prediction of Property				
Property	I	II	III	IV	V
Boiling temperature	0.43	0.16	0.04	0.24	0.67
Critical temperature	0.19	0.07	0.19	0.16	0.77
Critical pressure	0.23	1.00	2.49	5.93	9.33

Table 7. Structure–Structure Correlation Data and Coefficients for Butylcyclopentane, 1-Chlorohexane, and 1-Bromobutane

No.	Compound	C Atoms	W	Boiling Temp. (K)*	Coefficient
1 (target)	Butylcyclopentane	9	99	429.80	
2	Methylcyclopentane	6	26	344.96	0.443645
3	Ethylcyclopentane	7	43	376.62	−1.887290
4	Propylcyclopentane	8	67	404.11	2.443645
1 (target)	1-Chlorohexane	6	56	408**	
2	1-Chloropropane	3	10	319.67	1.794872
3	1-Chlorobutane	4	20	351.58	−4.589744
4	1-Chloropentane	5	35	381.54	3.794872
1 (target)	1-Bromobutane	4	20	374.75	
2	1-Bromoethane	2	4	311.5	−0.638767
3	1-Bromopropane	3	10	344.15	1.548458
4	1-Bromoheptane	7	84	452.05	0.090308

*Data from the DIPPR database.[16]
**Value from the NIST[17] database.

For the case of 1-bromobutane, all the pertinent data are available for both the target and the predictive compounds in the DIPPR database (experimental or predicted data). The prediction error for all the properties is well below the "reliability" of the data (<10% for P_c, for example).

To further verify those results we repeated all the calculations done with the Wiener index (shown in Table 7), using instead the normal boiling temperature as the noncollinear descriptor. No significant difference between the accuracy of the predictions was found.

It should be noted that when T_b is used as a descriptor, its experimental error also has to be considered when estimating the maximal error in the predicted property using the error propagation formula. The equations that apply in such a case are shown in the Appendix.

Long-range extrapolation

It was shown in the previous sections that the SC-QS2PR method yields predictions of most properties within experimental level when interpolation (narrow or wide range) or short-range extrapolation are involved. To test the applicability of the SC-QS2PR technique to property predictions, which involve extrapolation to high numbers of carbon atoms, we tested the possibility of predicting the boiling point temperatures of normal alkanes of 22 to 60 C atoms. The predictive compounds used are *n*-nonadecane, *n*-eicosane, and *n*-heneicosane. These were selected because they are the alkanes of the highest carbon number for which experimental boiling temperatures are available in the DIPPR database. The NIST[17] database provides four additional experimental values—up to C25—but these were used only to check the predictions.

The predicted T_b values are shown in Figure 1, where they are compared with values predicted by the methods of Marano and Holder,[2] Chickos,[9] and Brauner et al.[13] It can be seen that up to C36 the SC-QS2PR gives predictions that are consistent with the other two techniques. However, for C40 and higher numbers of C atoms, the trend indicated by the SC-QS2PR predictions is wrong, whereby the predicted T_b values decrease

Table 8. Percentage Error in Predicting Properties of Butylcyclopentane

			DIPPR Database[16]			SC-QS2PR Prediction	
No.	Property	Units	Reported Value	Data Type	Reliability	Value	Prediction Error (%)
1	Critical temperature	K	621	Predicted	<5%	617.93	0.49
2	Critical pressure	Pa	2.7E+06	Predicted	<10%	2.64E+06	2.78
3	Critical volume	m^{3} kmol^{-1}	0.483	Predicted	<25%	0.480	0.69
4	Crit compress factor	Unitless	0.254	Defined	None	0.251	1.09
5	Liq molar volume	m^{3} kmol^{-1}	0.161572	Experimental	<1%	0.1619681	0.25
6	IG heat of formation	J kmol^{-1}	−1.68E+08	Experimental	<3%	−1.69E+08	0.60
7	IG Gibbs of formation	J kmol^{-1}	6.2E+07	Defined	<3%	6.20E+07	0.25
8	IG absolute entropy	J kmol^{-1} K	4.54E+05	Experimental	<3%	4.5.+05	0.54
9	Std heat of formation	J kmol^{-1}	−2.14E+08	Experimental	<1%	−2.15E+08	0.67
10	Std Gibbs of formation	J kmol^{-1}	4.91E+07	Defined	<1%	4.94E+07	0.57
11	Std absolute entropy	J kmol^{-1} K	3.44E+05	Experimental	<1%	3.4.+05	0.71
12	Std net heat of comb	J kmol^{-1}	−5.50E+09	Experimental	<1%	−5.50E+09	0.00
13	Acentric factor	Unitless	0.37188	Defined	None	0.3899373	4.86
14	Radius of gyration	M	4.63E−10	Defined	<3%	4.541E−10	1.92
15	Solubility parameter	(J · m^{-3})$^{0.5}$	1.64E+04	Defined	<5%	1.65E+04	0.34
16	van der Waals volume	m^{3} kmol^{-1}	9.09E−02	Defined	<1%	9.08E−02	0.11
17	van der Waals area	m^{2}	1.16E+09	Defined	<1%	1.15E+09	0.68
18	Refractive index	Unitless	1.4293	Experimental	<0.2%	1.4288	0.03
19	Flash point	K	305	Predicted	<5%	305.2	0.07
20	Lower flamm. limit	vol % in air	0.8	Predicted	<10%	0.7778177	2.77
21	Upper flamm. limit	vol % in air	5.9	Predicted	<25%	6.6989209	13.54
22	Lower flamm. limit temp.	K	306			306.22542	0.07
23	Upper flamm. limit temp.	K	345			347	0.58

Table 9. Prediction Error in Predicting Properties of 1-Chlorohexane and 1-Bromobutane

			1-Chlorohexane			1-Bromobutane		
No.	Property	Unit	Reported Value*	Predicted Value	Prediction Error (%)	Reported Value**	Predicted Value	Prediction Error (%)
1	Critical temperature	K	594.6	597.7	0.53	577	579.34	0.41
2	Critical pressure	Pa		3.01E+06		4.54E+06	4.64E+06	2.22
3	Critical volume	$m^3\ kmol^{-1}$		0.404		0.319	0.318	0.31
4	Crit. compress factor	Unitless		0.246		0.302	0.311	2.98
5	Liq molar volume†	$m^3\ kmol^{-1}$		0.138		0.108	0.1077	0.28

*Value from the NIST database.[17]
**Data from the DIPPR database.[16]
†At 25°C.

with the carbon number rather than increase. This trend can be explained by inspection of Table 10, showing the structure–structure correlation parameters and the error multiplication factors ($\Sigma_i\ |\beta_i|$) associated with various extrapolation distances. The absolute values of the model parameters are significantly larger than those obtained with interpolation, and they increase with increasing extrapolation distance. Consequently, the error multiplication factor is already >7 for extrapolation to *n*-docosane, whereas this factor was <2 for all the cases that involved interpolation. The error multiplication factor grows very fast with extrapolation distance and reaches >1000 for C40. Although the factor provides only the upper limit of the propagated error, and the real error is probably much smaller, its increasing value can explain the large errors associated with long-range extrapolations.

The trends of the predictions provided by the other three methods are correct, with up to 4% differences between their respective predicted values. The method of Marano and Holder[2] yields the highest values and the method of Chickos[9] the lowest predictions, whereas the targeted QSPR method yields values in between these values. It should be pointed out that the full-scale QS2PR method[12] using the same predictive compounds (with 19, 20, and 21 C atoms) but 99 molecular descriptors, yields predictions that are very close to the values predicted by the other methods ($T_b = 804$ K for C40, $T_b = 830$ K for C44, and $T_b = 898$ K for C60). This demonstrates that the limited number of descriptors included in the SC-QS2PR model (Eq. 4) does not contain a sufficient amount of information to allow for long-range extrapolation. Identification of exactly what amount of structural information is needed for such tasks and how the amount of information relates to the range of extrapolation will be the object of our further studies.

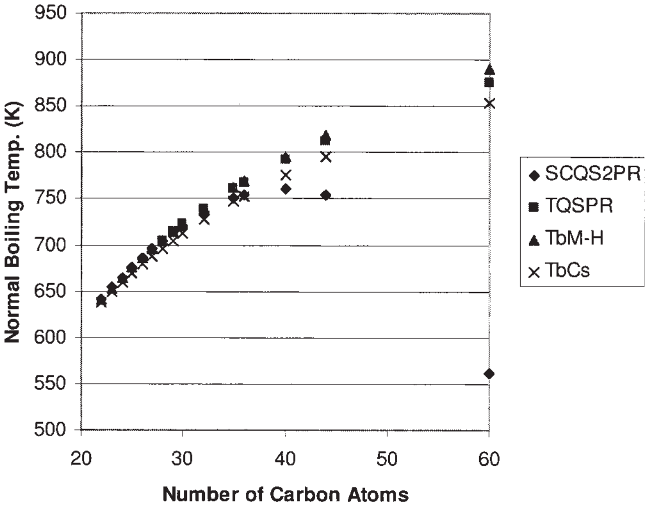

Figure 1. Predicting T_b for normal alkanes with high C numbers by SC-QS2PR extrapolation.

Comparison with other prediction methods: TQSPR (Brauner et al.[13]); TbM-H (Marano and Holder[2]); and TbCS (Chickos[9]).

Conclusions

It has been demonstrated that for homologous series, a simplified structure–structure relationship, based on three predictive compounds, can be derived using the number of carbon atoms and one additional noncollinear molecular descriptor. The so-obtained shortcut QS2PR model yields prediction within the experimental error level for liquid-phase and gas-phase properties if the correlation is used for interpolation or close-range extrapolation, as demonstrated with members of six homologous series, including hydrocarbons, oxygen, and halogene compounds.

Comparison with the existing QSPR,[11] QS2PR,[12] targeted QSPR,[13] and asymptotic behavior correlations[2] shows that when used for interpolation and short-range extrapolation the proposed method provides the same precision. However, because it is a QS2PR method, the coefficients of the structural correlation are used to predict all structure-dependent properties. It can still estimate the properties of the target when experimental data for only three (or even two) com-

Table 10. Structure–Structure Model Parameters and Error Multiplication Factors in Extrapolation

No. of C Atoms	Coefficient			Error Multiplication Factor: $\sum_i \|\beta_i\|$
	β_1	β_2	β_3	
22	1.05	−3.1	3.05	7.20
23	3.2	−8.4	6.2	17.80
24	6.5	−16	10.5	33.00
25	11	−26	16	53.00
26	16.75	−38.5	22.75	78.00
27	23.8	−53.6	30.8	108.20
28	32.2	−71.4	40.2	143.80
29	42	−92	51	185.00
30	53.25	−115.5	63.25	232.00
32	80.3	−171.6	92.3	344.20
35	133	−280	148	561.00
36	154	−323	170	647.00
40	256.5	−532	276.5	1065.00
44	391	−805	415	1611.00
60	1313	−2665	1353	5331.00

pounds in the series are available. In contrast to most existing techniques, the proposed method provides an upper limit of the prediction error using the experimental errors of the predictive compound data and the error propagation formula.

When a long-range extrapolation is involved, the amount of structural information used by the present SC-QS2PR method might be insufficient to provide reliable property prediction. To improve the prediction for long-range extrapolation, the number of predictive compounds and descriptors has to be increased. The more complex full-scale QS2PR method[12] and the targeted QSPR[13] method are also appropriate for long-range extrapolation. However, the authors intend to carry out future investigations in greater detail with respect to the balance between the structural variation of target compounds and the necessary number and information quality of the predictive compounds and descriptors.

Literature Cited

1. Prausnitz JM, Tavares FW. Thermodynamics of fluid-phase equilibria for standard chemical engineering operations. *AIChE J.* 2004; 50:739–761.
2. Marano JJ, Holder GD. General equations for correlating the thermophysical properties of *n*-paraffins, *n*-olefins and other homologous series 1 formalism for developing asymptotic behavior correlations. *Ind Eng Chem Res.* 1997;36:1887–1894.
3. Gao WZ, Robinson RL, Gasem KAM. Improved correlations for heavy *n*-paraffin physical properties. *Fluid Phase Equilib.* 2001; 179:207–216.
4. Nikitin ED, Pavlov PA, Bogatishcheva NS. Critical properties of long-chain substances from the hypothesis of functional self-similarity. *Fluid Phase Equilib.* 2005;235:18–23.
5. Wakeham WA, Cholakov GSt, Stateva RP. Liquid density and critical properties of hydrocarbons estimated from molecular structure. *J Chem Eng Data.* 2002;47:559–570.
6. Kontogeorgis GM, Tassios DP. Critical constants and acentric factors for long-chain alkanes suitable for corresponding states applications. A critical review. *Chem Eng J.* 1997;66:35–49.
7. Yan X, Dong Q, Hong H. Reliability analysis of group-contribution methods in predicting critical temperatures of organic compounds. *J Chem Eng Data.* 2003;48:374–380.
8. Coniglio L, Nouviaire A. A method for estimating the normal boiling point of heavy hydrocarbons suitable for a group-contribution-based equation of state. *Ind Eng Chem Res.* 2001;40:1781–1790.
9. Chickos JS. Hypothetical thermodynamic properties: The boiling and critical temperatures of polyethylene and polytetrafluoroethylene. *J Chem Eng Data.* 2004;49:518–526.
10. Nikitin ED, Popov AP, Bogatishcheva NS. Critical point measurements for five *n*-alkylcyclohexanes (C6 to C10) by the pulse-heating method. *J Chem Eng Data.* 2003;48:1137–1140.
11. Cholakov GSt, Wakeham WA, Stateva RP. Estimation of normal boiling temperature of industrially important hydrocarbons from descriptors of molecular structure. *Fluid Phase Equilib.* 1999;163:21–42.
12. Shacham M, Brauner N, Cholakov GSt, Stateva RP. Property prediction by correlations based on similarity of molecular structures. *AIChE J.* 2004;50:2481–2492.
13. Brauner N, Stateva RP, Cholakov GSt, Shacham M. A structurally "targeted" QSPR method for property prediction. *Ind Eng Chem Res.* 2006;45:000–000 (DOI: 10.1021/ie051155o).
14. Shacham M, Brauner N. The SROV program for data analysis and regression model identification. *Comput Chem Eng.* 2003;27:701–714.
15. Brauner N, Shacham M, Cholakov GSt, Stateva RP. Property prediction by similarity of molecular structures—Practical application and consistency analysis. *Chem Eng Sci.* 2005;60:5458–5471.
16. Rowley RL, Wilding WV, Oscarson JL, Yang Y, Zundel NA. *DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties.* Provo, UT: Brigham Young University; 2006 (http//dippr.byu.edu).
17. National Institute of Standards and Technology (NIST). In: Linstrom PJ, Mallard WG, eds.*Chemistry WebBook, NIST Standard Reference Database Number 69*. Gaithersburg, MD:NIST; June 2005 (http://webbook.nist.gov).
18. Labanowski JK, Motoc I, Dammkoehler RA. The physical meaning of topological indices. *Comput Chem.* 1991;15:47–53.
19. Godavarthy SS, Robinson RL Jr, Gasem KAM. An improved structure-property model for predicting melting-point temperatures. *Ind Eng Chem Res.* 2006;45:5117–5126.
20. Zbogar AZ, Da Silva Lopes FV, Kontogeorgis GM. Approach suitable for screening estimation methods for critical properties of heavy compounds. *Ind Eng Chem Res.* 2006;45:476–480.

Appendix: Estimating Prediction Error When Experimental Values Are Used as Descriptors

In case property values (such as boiling point), which are subjected to experimental error, are used to calculate the coefficients of the "structure–structure" relation (replacing the Wiener index, for example), the maximal error in the predicted property value (Δy_p) should also account for the additional error introduced in the calculated coefficients values, β_i. In this case, an estimate of error in the predicted property value is given by

$$\Delta y_p = \sum_j |\beta_j| \Delta x_{pj} + \sum_j x_{pj} \Delta\beta_j \tag{A1}$$

For three predictive compounds, $y_p = \Sigma_{j=1}^{3} \beta_j x_{pj}$, and the Eq. 4a system are used to calculate the coefficient values. The maximal errors in the parameter values arising from errors in the values of y_C and x_{Ci} (Δy_C and Δx_{Ci}, respectively) are then given by

$$\begin{aligned}
\Delta\beta_1 &= \frac{(n_{p3} - n_{p2})}{(n_{p2} - n_{p1})} \Delta\beta_3 \\
\Delta\beta_2 &= -\frac{(n_{p3} - n_{p1})}{(n_{p2} - n_{p1})} \Delta\beta_3 \\
\Delta\beta_3 &= \left| \frac{\partial\beta_3}{\partial y_C} \right| \Delta y_C + \sum_{i=1}^{3} \left| \frac{\partial\beta_3}{\partial x_{Ci}} \right| \Delta x_{Ci}
\end{aligned} \tag{A2}$$

with

$$\frac{\partial\beta_3}{\partial y_C} = \frac{(n_{p2} - n_{p1})}{(x_{C3} - x_{C1})(n_{p2} - n_{p1}) - (n_{p3} - n_{p1})(x_{C2} - x_{C1})}$$

$$\frac{\partial\beta_3}{\partial x_{C3}} = \frac{-(n_{p2} - n_{p1})}{(x_{C3} - x_{C1})(n_{p2} - n_{p1}) - (n_{p3} - n_{p1})(x_{C2} - x_{C1})} \beta_3$$

$$\frac{\partial\beta_3}{\partial x_{C1}} = \frac{(n_1 - n_{p2}) - (n_{p3} - n_{p2})\beta_3}{(x_{C3} - x_{C1})(n_{p2} - n_{p1}) - (n_{p3} - n_{p1})(x_{C2} - x_{C1})}$$

$$\frac{\partial\beta_3}{\partial x_{C2}} = \frac{(n_{p1} - n_1) - (n_{p3} - n_{p1})\beta_3}{(x_{C3} - x_{C1})(n_{p2} - n_{p1}) - (n_{p3} - n_{p1})(x_{C2} - x_{C1})}$$

Manuscript received Mar. 21, 2006, and revision received Oct. 18, 2006.